Genetics Meets Population Science: Understanding the Importance of Response Rates in Genetic Discovery and Prediction in the UK Biobank

Xiaoyuan Zhong[1], Yuchang Wu[2], Yunong Lin[3], Lauren L. Schmitz[4,5], Qiongshi Lu[2,3,5], Jason M. Fletcher[4,5,6]

1 University of Wisconsin–Madison
2 Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison
3 Department of Statistics, University of Wisconsin–Madison
4 La Follette School of Public Affairs, University of Wisconsin–Madison
5 Center for Demography of Health and Aging, University of Wisconsin–Madison
6 Department of Sociology, University of Wisconsin–Madison

Background

- UK Biobank (UKB) invited about 9,000,000 people in order to obtain 500,000 respondents, which means the participation rate is merely above 5%. Also, there are systematic differences between participants and invitees[1] (***Figure S1*** and ***Table S1***).
- Participation bias due to nonresponse could contaminate the results of genetic studies[2]. Here, we propose a framework to estimate the liability of participation and use the estimated liability to adjust for participation bias in UKB.

Methods

- We utilized the marginal distributions of age at recruitment, sex, region of residence, and Townsend index among both UKB invitees and participants to estimate the liability of participation (***Figure 1***).
- We calculated participation probability for each UKB individual based on estimated liability. Sampling weight was constructed as the inverse of participation probability.
- We used the estimated liability to adjust for participation bias in GWAS via Heckman correction[3].

Data

- Sampling weights were used to predict participation in four optional components[4] of UKB baseline study.
- Heckman correction was applied on self-conducted GWAS of educational attainment (EA).

Results

- Our sampling weights can predict UKB optional components participation status (***Table 1***).
- We saw a very minor reduction in heritability after applying Heckman correction on EA GWAS (***Figure 2*** and ***Figure 3***).

Future directions

- Apply Heckman correction on more GWAS, check the pattern for different traits.
- Look into other approaches to justify and improve our participation liability estimation.

Estimate participation liability from marginal distributions of covariates

Figure 1. We modeled the participation liability Z^* on the entire sampling frame with a linear equation. W is a centered design matrix containing k covariates and ε is a normal random error. The multiple linear regression estimator of γ, $\hat{\gamma}$, can be inferred from marginal linear regression estimator, $\tilde{\gamma}$. $\tilde{\gamma}$ is a function of the marginal logistic regression estimator, $\tilde{\delta}$, and we can directly calculate $\tilde{\delta}$ from data (see ***Figure S2***).

$$\boldsymbol{Z^* = W\gamma + \varepsilon}$$
$$\uparrow$$
$$\boldsymbol{\gamma \approx \hat{\gamma} = (W^T W)^{-1} W^T Z^*}$$
$$\uparrow$$
$$\boldsymbol{\hat{\gamma} = (W^T W)^{-1} \begin{pmatrix} w_1^T w_1 & & \\ & \ddots & \\ & & w_k^T w_k \end{pmatrix} \tilde{\gamma}}$$
$$\uparrow$$
$$\boldsymbol{\tilde{\gamma}_j = (W_j^T W_j)^{-1} W_j^T Z^* \approx f(\tilde{\delta}_j)}$$
$$\uparrow$$
$$\boldsymbol{\tilde{\delta}_j = \log(OR_j)}$$

Sampling weights can predict participation in UKB optional components

	Estimate	SE	P-value	AUC	N
AM	-0.106	0.003	2.042E-217	0.549	407727
FFQ	-0.042	0.004	1.814E-26	0.534	270926
MHQ	-0.044	0.004	9.688E-25	0.527	240933
PAS	-0.027	0.005	2.974E-08	0.521	191668

Table 1. We ran logistic regression of option components participation status on sampling weights. The direction of effect size estimate aligns with our expectation, as a larger sampling weight should imply lower participation likelihood. P-values are all significant, and AUC are all above 50%. AM: aide memoire, FFQ: food frequency questionnaire, MHQ: mental health questionnaire, PAS: physical activity study.

Heckman correction on EA GWAS

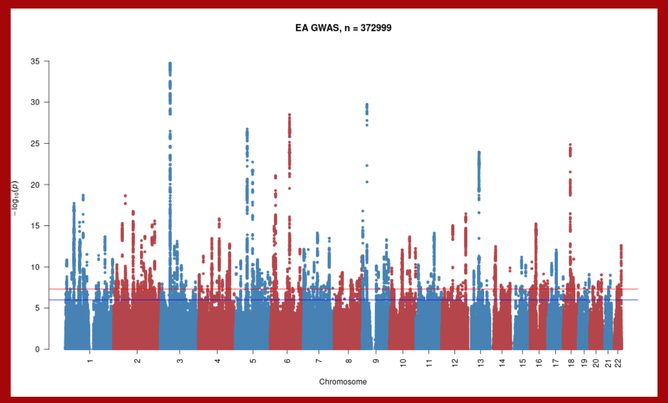

Figure 2. Manhattan plot for regular EA GWAS. Phenotype definition follows from the EA3 GWAS[5]; covariates include age, sex, batch, and first 20 genetic principal components. h^2=0.1466.

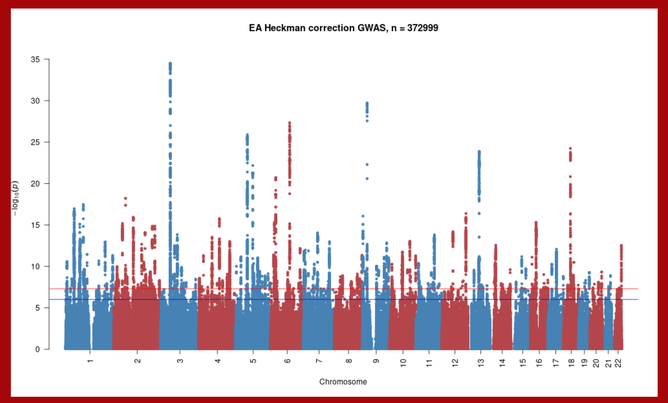

Figure 3. Manhattan plot for EA GWAS after Heckman correction. Adjusted for the same set of covariates as in regular EA GWAS. h^2=0.1406, r_g with regular EA GWAS is 0.9972.

Supplementary Material

Heckman selection model

GWAS model:

$$\boldsymbol{Y = X\beta + \varepsilon}$$

Participation liability model:

$$\boldsymbol{Z^* = W\gamma + \tau}$$

A binary variable $\boldsymbol{Z}$ represents whether a population unit is in the sample:

$$Z_i = 1_{\{W_i\gamma + \tau_i > T\}}$$

Suppose ε_i and τ_i follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \tau_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho \\ \rho & \sigma_\tau^2 \end{pmatrix}\right)$$

Bias is a function of liability:

$$E(Y_i | \boldsymbol{X_i}, Z_i = 1, \boldsymbol{W_i})$$
$$= \boldsymbol{X_i\beta} + \rho\frac{\sigma_\varepsilon}{\sigma_\tau} E(\tau_i | \tau_i > -\boldsymbol{W_i\gamma} + T, \boldsymbol{W_i})$$
$$= \boldsymbol{X_i\beta} + \rho\sigma_\varepsilon \frac{\phi\left(\frac{W_i\gamma + T}{\sigma_\tau}\right)}{\Phi\left(\frac{W_i\gamma + T}{\sigma_\tau}\right)} \quad \text{bias}$$

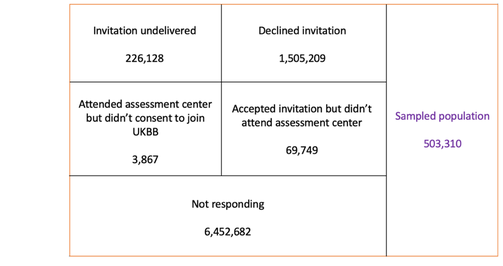

Figure S1. Sampling structure of the UKB. The frame population contains about 9,000,000 individuals, but due to various reasons only 500,000 of them joined the study.

	UK Biobank Invitees (n = 8,761,869) No. (%)	UK Biobank Participants[b] (n = 503,310) No. (%)
Sex		
Men	4,468,580 (51.0)	229,486 (45.6)
Women	4,293,289 (49.0)	273,824 (54.4)
Age group at time of invitation, years		
40–44	1,770,821 (20.2)	53,953 (10.7)
45–49	1,780,661 (20.3)	66,438 (13.2)
50–54	1,506,431 (17.2)	76,808 (15.3)
55–59	1,366,076 (15.6)	91,953 (18.3)
60–64	1,323,219 (15.1)	121,419 (24.1)
65–70	1,014,661 (11.6)	92,739 (18.4)
Region		
South East	717,053 (8.2)	50,498 (10.0)
London	1,330,405 (15.2)	61,982 (12.3)
East Midlands	618,724 (7.1)	40,399 (8.0)
West Midlands	877,480 (10.0)	41,024 (8.2)
South West	440,989 (5.0)	42,340 (8.4)
North West	1,758,685 (20.1)	82,386 (16.4)
Yorkshire and the Humber	1,074,321 (12.3)	65,377 (13.0)
North East	957,005 (10.9)	61,714 (12.3)
Scotland West	435,661 (5.0)	18,576 (3.7)
Scotland East	211,357 (2.4)	17,309 (3.4)
Wales	340,189 (3.9)	21,705 (4.3)
Townsend deprivation score[c]		
Less deprived (<-2)	3,137,375 (35.8)	260,185 (51.8)
Average (≥-2 to <2)	2,971,606 (34.0)	159,483 (31.8)
More deprived (≥2)	2,645,024 (30.2)	82,360 (16.4)

Table S1. Distributions of various features among UKB invitees and participants. Adapted from Web Table 2, Fry et al. (2017).

References

[1] Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
[2] Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
[3] Heckman, J. Sample bias as a specification error. *Econometrica* **47**, 153–162 (1979).
[4] Tyrrell, J. et al. Genetic predictors of participation in optional components of UK Biobank. *bioRxiv* (2020).
[5] Lee, J.J., Wedow, R., Okbay, A. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50,** 1112–1121 (2018).

About me

Xiaoyuan Zhong is an undergraduate student at University of Wisconsin–Madison. His research interests lie in the area of statistical genetics and genetic epidemiology.

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

LU LABORATORY
statistical genetics & genome informatics